frontiers in Oncology

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/fonc.2021.781798

Automatic Sequence-Based Network for Lung Diseases Detection in Chest CT

Jinkui Hao[1,2†], Jianyang Xie[1†], Ri Liu[3†], Huaying Hao[1], Yuhui Ma[1,2], Kun Yan[3], Ruirui Liu[4], Yalin Zheng[5], Jianjun Zheng[4], Jiang Liu[1,6], Jingfeng Zhang[3] and Yitian Zhao[1,7,8*]*

[1] Cixi Institute of Biomedical Engineering, Ningbo Institute of Material Technology and Engineering, Chinese Academy of Sciences, Ningbo, China, [2] School of Optical Technology, University of Chinese Academy of Sciences, Beijing, China, [3] Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, China, [4] School of Medicine, Ningbo University, Ningbo, China, [5] Department of Eye and Vision Science, University of Liverpool, Liverpool, United Kingdom, [6] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China, [7] Zhejiang International Scientific and Technological Cooperative Base of Biomedical Materials and Technology, Ningbo Institute of Material Technology and Engineering, Chinese Academy of Sciences, Ningbo, China, [8] Zhejiang Engineering Research Center for Biomedical Materials, Ningbo Institute of Material Technology and Engineering, Chinese Academy of Sciences, Ningbo, China

OPEN ACCESS

Edited by:
Guang Yang,
Imperial College London,
United Kingdom

Reviewed by:
Anand Nayyar,
Duy Tan University, Vietnam
Seyedali Mirjalili,
Torrens University Australia, Australia
Tao Zhou,
Nanjing University of Science and Technology, China
Zhili Chen,
Shenyang Jianzhu University, China

***Correspondence:**
Yitian Zhao
yitian.zhao@nimte.ac.cn
Jingfeng Zhang
jingfeng.zhang@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Cancer Imaging and Image-directed Interventions, a section of the journal Frontiers in Oncology

Received: *23 September 2021*
Accepted: *01 November 2021*
Published: *02 December 2021*

Citation:
Hao J, Xie J, Liu R, Hao H, Ma Y, Yan K, Liu R, Zheng Y, Zheng J, Liu J, Zhang J and Zhao Y (2021) Automatic Sequence-Based Network for Lung Diseases Detection in Chest CT. Front. Oncol. 11:781798. doi: 10.3389/fonc.2021.781798

Objective: To develop an accurate and rapid computed tomography (CT)-based interpretable AI system for the diagnosis of lung diseases.

Background: Most existing AI systems only focus on viral pneumonia (e.g., COVID-19), specifically, ignoring other similar lung diseases: e.g., bacterial pneumonia (BP), which should also be detected during CT screening. In this paper, we propose a unified sequence-based pneumonia classification network, called SLP-Net, which utilizes consecutiveness information for the differential diagnosis of viral pneumonia (VP), BP, and normal control cases from chest CT volumes.

Methods: Considering consecutive images of a CT volume as a time sequence input, compared with previous 2D slice-based or 3D volume-based methods, our SLP-Net can effectively use the spatial information and does not need a large amount of training data to avoid overfitting. Specifically, sequential convolutional neural networks (CNNs) with multi-scale receptive fields are first utilized to extract a set of higher-level representations, which are then fed into a convolutional long short-term memory (ConvLSTM) module to construct axial dimensional feature maps. A novel adaptive-weighted cross-entropy loss (ACE) is introduced to optimize the output of the SLP-Net with a view to ensuring that as many valid features from the previous images as possible are encoded into the later CT image. In addition, we employ sequence attention maps for auxiliary classification to enhance the confidence level of the results and produce a case-level prediction.

Results: For evaluation, we constructed a dataset of 258 chest CT volumes with 153 VP, 42 BP, and 63 normal control cases, for a total of 43,421 slices. We implemented a comprehensive comparison between our SLP-Net and several state-of-the-art methods across the dataset. Our proposed method obtained significant performance without a

large amount of data, outperformed other slice-based and volume-based approaches. The superior evaluation performance achieved in the classification experiments demonstrated the ability of our model in the differential diagnosis of VP, BP and normal cases.

Keywords: deep learning, CT, CNN, ConvLSTM, lung diseases

1 INTRODUCTION

COVID-19, the latest in viral pneumonia diseases, is an acute respiratory syndrome that has spread rapidly around the world since the end of 2019, having a devastating effect on the health and well-being of the global population (1, 2). To diagnose viral pneumonia (limited to COVID-19 in our work), reverse transcription-polymerase chain reaction (RT-PCR) has widely been accepted as the gold standard. However, shortages of equipment and strict requirements for testing environments limit the rapid and accurate screening of suspected subjects. Furthermore, RT-PCR testing is also reported to suffer from a high false-negative rate (3), with a low sensitivity of only 71%. In clinical practice, radiological imaging techniques, e.g., X-rays and computed tomography (CT), have also been demonstrated to be effective in diagnosis, and also follow-up assessment and evaluation of disease evolution (4, 5). CT is the most widely used imaging technique, due to its high resolution and three-dimensional (3D) view, and its relatively high detection sensitivity of around 98% (6). For example, the study (5) found that the dynamic lesion process of viral pneumonia (from ground-glass opacity in the early stage to pulmonary consolidation in the late stage) can be observed in CT scans, and its CT manifestations have been emphasized.

Bacterial and viral pathogens are the two leading causes of pneumonia, but require very different forms of management (7). Bacterial pneumonia requires urgent referral for immediate antibiotic treatment, while viral pneumonia is treated with supportive care. Therefore, accurate classification of different types of pneumonia is imperative for timely diagnosis and treatment. However, the imaging features of viral and bacterial infections are not often compared, and the only imaging feature that was significantly different between the viral and bacterial lung infection was the frequency of diffuse airspace disease (8). In the case of a typical viral pneumonia, in clinical practice, it is difficult to accurately differentiate viral pneumonia from bacterial pneumonia. See **Figure 1** as an example. In clinical practice, especially in primary medical institutions, the consistency of imaging diagnosis of pneumonia pathogens is poor (9–11). Moreover, it is time consuming for radiologists to read CT volumes that contain hundreds of 2D slices. As such, it is of great practical significance to quickly and accurately identify pathogens to guide individualized anti-infectious treatment and minimize and delay the occurrence of drug resistance.

As an emerging technology in medical image analysis, artificial intelligence (AI) has been widely employed for lesion segmentation, and for clinical assessment and diagnosis of lung-related diseases *via* radiological imaging (12, 13). Recently, many novel AI techniques for viral pneumonia have been presented (1). For instance, Ouyang et al. (14) proposed a dual-sampling attention network for the differential diagnosis of COVID-19 from Community Acquired Pneumonia (CAP) (14, 15), and Fan et al. (16) introduced an automatic COVID-19 lung infection lesion segmentation method using a deep network. These works

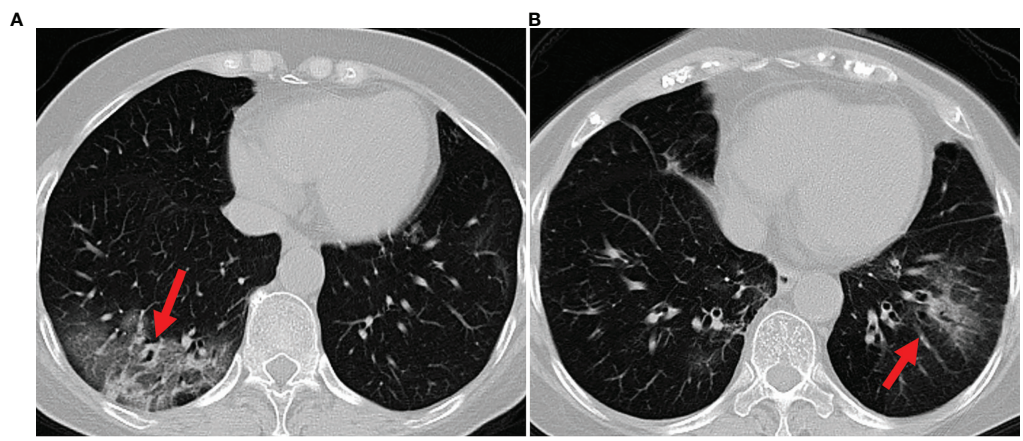

FIGURE 1 | Example axial CT slices of viral pneumonia **(A)**, bacterial pneumonia **(B)**. Accurate classification of different types of pneumonia is imperative for timely diagnosis and treatment. However, viral pneumonia and bacterial pneumonia display similar appearances in a CT image, which makes it difficult to accurately differentiate a patient with viral pneumonia from a case of bacterial pneumonia.

are useful for detecting and controlling of the spread of COVID-19. However, there are very few studies on differentiating COVID-19 from other etiological pneumonias, despite success in using deep learning (DL) approaches to discriminate bacterial and viral pneumonias in pediatric chest radiographs (17, 18).[1]

Further, most existing works make use of 2D CT slices, and the lack of continuity information makes it impossible to capture the true spatial distribution of the lesion in the lungs. To this end, some recent studies have attempted to use entire 3D volumes to train a 3D classification or segmentation model directly (14, 19) , achieving a slightly better performance than the approaches based on 2D slices. However, these 3D volume based approaches greatly increase the computational load, and require much more powerful and expensive hardware configurations. Additionally, 3D volumes may contain large portions of redundant information, which leads to great difficulty in accurately identifying small lesions. The imaging appearance of viral and bacterial lung infection has considerable similarity, and that, in any individual case, the viral pneumonia cannot reliably be distinguished from bacterial infections (8, 20). For example, viral pneumonia and bacterial pneumonia have some image features in common, such as ground glass opacities and interstitial changes in the peripheral zone of lungs, and accompanied by partial consolidation. The only imaging feature that was significantly different between the viral and bacterial lung infection was the frequency of diffuse airspace disease (8). Precise characterization of the spatial morphology of the infected regional lesions is essential to distinguish the two infection types by CT imaging.

In this paper, we treat the spatially continuous 2D CT slices as a time sequence and proposed a unified sequence-based pneumonia classification network (SLP-Net) for differentiating viral pneumonia (VP) from bacterial pneumonia (BP) and normal control cases. Our network comprises a CNN encoder and the ConvLSTM module, and sequence attention maps are used for auxiliary classification. As stated above, the precise characterization of the lesion is the key to distinguish the different pneumonia types. The combination of these components ensures the model pay more attention to spatial morphology of the lesion during the decision making. Specifically, the encoder with multi-scale receptive fields is first used to extract local representations of the sequence. Then we apply the ConvLSTM to acquire spatial information of these sequence features, modeling the distribution of the lesion. To optimize the SLP-Net, we introduce a novel adaptive-weighted cross-entropy (ACE) loss, with a view to ensuring that as many valid features from the previous images as possible are encoded into the subsequent CT image. Given the fact that the final diagnosis conclusion needs to be made for each patient, case-based prediction rather than a slice- or sequence-based prediction is more valuable. To obtain case-based prediction, in addition to the classification result of the sequence, we also use sequence attention maps to aid the case-level classification, aiming to enhance the confidence of the results. We collect a dataset of 258 chest CT volumes (153 VP, 42 BP, and 63 normal control cases). The experimental results show that the proposed SLP-Net achieves an accurate classification performance of viral pneumonia, bacterial pneumonia, and normal control, which could benefit the large-scale screening and control of viral pneumonia, and also enable efficient treatment for different types of pneumonia.

We organize the remainder of this paper as follows. In Section 2, the existing methods of AI-enpowered viral pneumonia analysis are briefly reviewed. In Section 3 we give detailed descriptions of collected datasets. Section 4 introduces the proposed SLP-Net. In Section 5, we present the experimental results and discuss the effectiveness, robustness, and efficiency of the SLP-Net. Section 6 concludes the paper and indicates directions for future work.

2 RELATED WORK

AI-based medical image analysis plays an essential role in the global fight against COVID-19, and a considerable number of approaches have been proposed in the past five months. This body of work on COVID-19 has focused primarily on two problems: lesion segmentation (16, 21, 22), and automated screening (23–31). For example (16), recently introduced a parallel partial decoder to aggregate high-level features, using an implicit reverse attention and explicit edge-attention to model boundaries and enhance representations so as to identify infected regions from 2D chest CT slices. To alleviate the shortage of labeled data, a semi-supervised segmentation framework based on a randomly selected propagation strategy was applied by (21). They proposed a relational approach, in which a non-local neural network module was introduced to efficiently learn both visual and geometric relationships among all convolutional features.

However, automated viral pneumonia (e.g., COVID-19) screening has attracted even more attention. For instance (32), introduced a COVID-19 detection method with multi-task DL approaches, using an inception residual recurrent convolutional neural network (CNN) with transfer learning. Their detection model achieved 84.67% accuracy from X-ray images (33). proposed a deep features fusion and ranking technique to detect COVID-19 in its early phase. They employed a pre-trained CNN structure to obtain a set of features, which were subsequently fused and evaluated with a support vector machine (SVM) classifier. In the classification task of COVID-19 and no COVID-19, their proposed method obtained 98.27% accuracy on their own dataset (34). applied a modified residual network, called DeepPneumonia, based on ResNet50 for slice-level classification, and could discriminate the COVID-19 patients from the bacteria pneumonia patients with an AUC of 0.95 (23). built multiple deep convolutional neural models for classifying chest X-ray images into normal and COVID-19 cases, which obtained 96.1% accuracy (35). proposed a unified latent representation to explore multiple features describing CT images from different views, a method that can completely encode information from different features aspects and is endowed with a promising class structure for separability. Performance in diagnosis for COVID-19 and community-

[1] https://github.com/HzFu/COVID19.

acquired pneumonia (CAP) is 95.5% in terms of accuarcy. An infection size-aware random forest method was introduced by (15) for the differentiation of COVID-19 from CAP, in which patients were automatically categorized into groups with different extensions of infected lesion sizes, followed by generation of random forests with each group for classification. The method achieved an accuracy of 89.4% in discriminating COVID-19 from CAP.

However, all of the above mentioned works are based on 2D images, the spatial correlation between consecutive CT scans is neglected by most slice-based methods, despite this being essential for the screening of lung diseases. A variety of volume-based methods have been proposed in an attempt to address this deficiency (19). proposed an attention-based deep 3D multi-instance learning method to screen COVID-19 from 3D chest CT sacns, using a weakly supervised learning framework that incorporates an attention mechanism into deep multi-instance learning, achieving an accuracy of 97.9% (14). proposed a 3D CNN to diagnose COVID-19 from CAP, in which a novel online attention module is combined with a dual-sampling strategy. The online attention module focuses on the infected regions when making diagnostic decisions. The dual-sampling strategy mitigates the imbalanced distribution in the sizes of infected regions between COVID-19 and CAP. Their method was evaluated in private dataset and achieved an accuracy of 87.5%. These 3D volume based approaches greatly increase the computational load, and require much more powerful and expensive hardware configurations. Additionally, 3D volumes may contain large portions of redundant information, which leads to great difficulty in accurately identifying small lesions. Overall, 2D slice based methods cannot take advantage of spatial continuity information and 3D volume based methods require much more expensive hardware configurations. In order to take advantage of the complementary information of 2D slices and 3D volumes, we treat the spatially continuous 2D CT slices as a time sequence, and divide the volume into multiple different temporal sequences of consecutive slices as the input.

3 MATERIALS AND METHODS

3.1 Materials

A total of 258 subjects were enrolled into this study, with 258 CT volumes, corresponding to 43,421 slices. Of the 258 subjects, 42 patients were confirmed positive for BP by clinical diagnosis (age: 59.5 ± 27.2; male/female: 36/6), 153 patients were positive for VP, confirmed by RT-PCR (age: 52.3 ± 12.7; male/female: 68/85), and 63 were control subjects (age: 35.8 ± 11.7; male/female: 33/30). The CT volumes of normal and VP patients were captured between January 29, 2020 and February 18, 2020, and the BP data was collected between January 2, 2019 and February 19, 2020. There is no statistically significant difference between the ages of the VP and BP subjects (P = >0.05), but both groups are significantly older than patients in the normal group (P < 0.001). CT examinations of all the enrolled patients were performed on a ScintCare CT16 (Minfound Inc, China) with standard chest imaging protocols. All the patients underwent CT scans during the end-inspiration without the administration of contrast material. Related parameters for chest CT scanning were listed as follows: field of view (FOV), 360 mm; tube voltage, 120 kV; tube current, 240 mA; helical mode; slice thickness, 5 mm; pitch, 1.5; collimation 16 × 1.2 mm; gantry rotation speed, 0.5 s/r; matrix, 512 × 512; software version, syngo CT 2014A; mediastinal window: window width of 350 HU, with a window level of 40 HU; and lung window: window width of 1,300 HU, with a window level of −500 HU.

CT volumes were retrospectively collected according to the history of laboratory investigations (e.g., sputum culture and reverse transcription-polymerase chain reaction), which we can generate the case-level labels. Meanwhile, professional radiologists (from the Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, China.) picked out the slice containing the infected region in each volume for the subsequent automatic generation of sequence labels: each volume was be divided into overlapping sequences containing n slices, with k overlapping slices between two sequences. If a sequence from VP volume contains the infected slice(s), the label for that sequence is 1; if it is from BP volume and contains infected slice(s), the label is 2; the label from normal volume is 0. It is worth noting that k and n are both hyperparameters, and in *Sensitivities to Hyperparameters* we discuss how to choose the values of these, as well as their impact on the classification results.

The sequences generated from the volume of VP and BP were not used for training if they did not contain any slices with lesion regions. If all normal slices from patients are used for training instead of excluding them, it will increase the proportion of normal control samples in the training set.The data imbalance may cause the model to over-fit the normal samples and tend to predict the samples with the lesion as normal. To avoid this, we exclude the normal slice from patients. For training and evaluation of the proposed method, as shown in **Table 1**, we split 258 volumes into 168 volumes (95 VP, 30 BP, and 43 normal controls) for training and 90 (58 VP, 12 BP, and 20 normal controls) volumes for testing.

3.2 Proposed Method

The architecture of our SLP-Net is shown in **Figure 2**, which consists of two main components: sequence CNNs, and ConvLSTM. The sequence CNNs with multi-scale receptive fields are employed to extract more discriminative high-level features from the CT sequence, while the ConvLSTM captures the axial dimensional dynamics of features. In addition, a sequence attention map is utilized as an auxiliary means to integrate the output of the network and obtain prediction results with a higher level of confidence. Finally, an adaptive-weighted cross-entropy (ACE) loss is used to optimize the whole model.

3.2.1 Sequence CNN With Multi-Scale Receptive Fields

A typical CNN model consists of a stack of convolution layers, interleaved with non-linear downsampling operations (e.g., max pooling) and point-wise nonlinearities (e.g., ReLU). The residual

TABLE 1 | Characteristics of training and testing CT dataset for identifying viral pneumonia (VP) from bacterial pneumonia (BP) and normal controls.

Cohort	VP		BP		Normal controls		Total	
	Volumes	Slices	Volumes	Slices	Volumes	Slices	Volumes	Slices
Training set	95	15,931	30	5,068	43	7,310	168	28,309
Testing set	58	7,832	12	3,107	20	4,173	90	15,112
Total	153	23,763	42	8,175	63	11,483	258	43,421

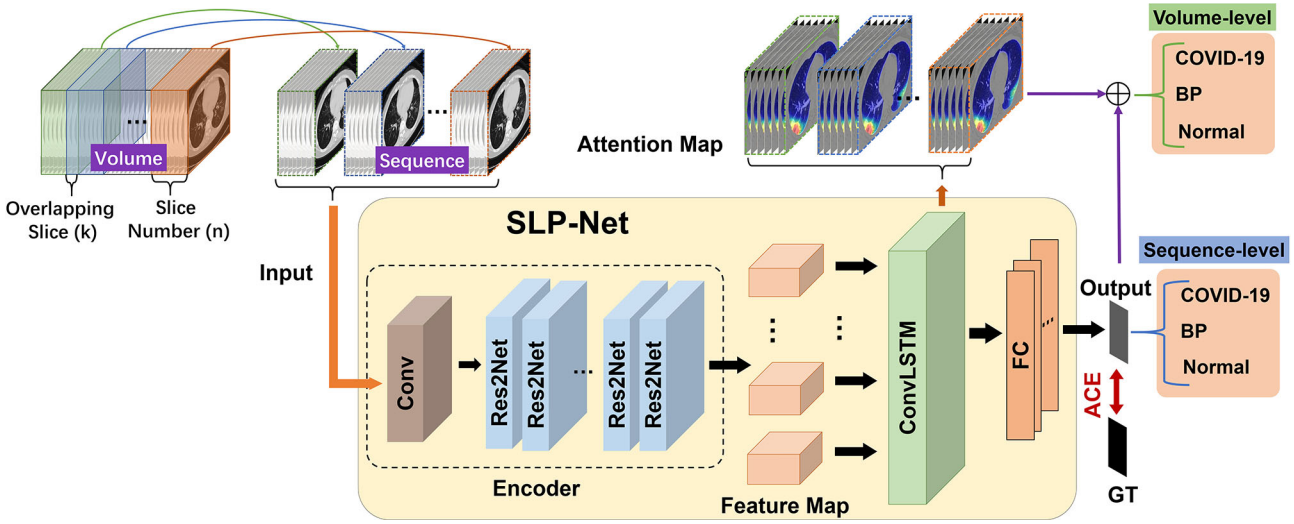

FIGURE 2 | Flowchart of our whole system for differentiating between viral pneumonia and bacterial pneumonia in a chest CT volume. Each volume is divided into overlapping sequences containing *n* slices during the training phase, such that the overlapping slices between two sequences are *k*. When predicting each volume during the testing phase, in addition to using the model to obtain the classification results of the sequence, we also introduce sequence attention maps for auxiliary classification to enhance the confidence level of the results. GT, ground truth; ACE, adaptive-weighted cross-entropy loss; FC, fully connected layer.

shortcut used in ResNet can reduce the over-fitting of the model, so that the depth of the network can be greater and achieve better performance. Taking into consideration the problems of overfitting and parameter cost, we employed ResNet (36) as our encoder backbone. The first four feature-extracting blocks are retained, without the average-pooling layer and the fully-connected layers.

Since VP and BP reveal similar appearances in CT images, we aim to obtain more discriminative features by employing multi-scale information, in order to distinguish them more accurately. Unlike most existing methods (37–39) that improve multi-scale ability by utilizing features with different resolutions, we apply a recently proposed multi-scale receptive fields technique (40) to enhance representation ability at a more granular level. Specifically, we apply a modified bottleneck with multi-scale ability, the Res2Net module, to replace a group of 3 × 3 filters used in the original bottleneck block of ResNet. As shown in **Figure 3**, after the 1 × 1 convolution, feature maps are split into *s* feature map subsets, denoted by x_i. Then, apart from x_1, each x_i goes through a corresponding 3 × 3 convolutional operator, denoted by $K_i(\cdot)$, where y_i is the output of $K_i(\cdot)$. The output of $K_{i-1}(\cdot)$is added to x_i, then sent to the next group of filters $K_i(\cdot)$. Thus, y_i can be defined as follows:

$$y_i = \begin{cases} x_i & i = 1; \\ K_i(x_i) & i = 2; \\ K_i(x_i + y_{i-1}) & 2 < i \le s. \end{cases} \quad (1)$$

In order to better integrate the information from different scales, all outputs y_i, where $i \in \{1,2, \ldots, s\}$, are concatenated and passed through a 1×1 convolution. Such splitting and concatenation strategies can force the convolution to process features more efficiently. Note that each 3×3 convolution operation $K_i(\cdot)$receives information from all the feature splits $\{x_j, j \le i\}$. Each time x_j performs a 3×3 convolution, the size of the receptive field will increase. Due to the combinatorial effect, the output of the Res2Net block contains different combinations of receptive field sizes/scales.

3.2.2 ConvLSTM With ACE Loss

Although the conventional fully-connected LSTM (FC-LSTM) can handle sequences of any length and capture long-term dependencies (41), it contains too much redundancy for spatial data, which is a critical problem for image sequences. Inspired by video object detection (42), we apply ConvLSTM (43) to process the feature sequences from the encoder.

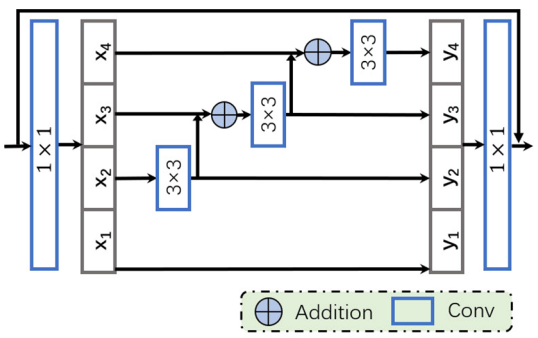

FIGURE 3 | A Res2Net module is utilized to extract more discriminative features.

As the convolutional counterpart of the FC-LSTM, the ConvLSTM introduces the convolution operation into the input-to-state and state-to-state transitions. The ConvLSTM can model axial dimensional dependencies while preserving spatial information. As with the FC-LSTM, the ConvLSTM unit (see **Figure 4**) includes an input gate i_t, a memory cell C_t, a forget gate f_t and an output gate o_t. The memory cell C_t, acting as an accumulator of the state information, is accessed, updated and cleared through self-parameterized controlling gates: i_t, o_t, and f_t. If the input gate is switched on, the new data is accumulated into the memory cell once an input arrives. Similarly, the past cell status C_{t-1} will be forgotten if the forget gate f_t is activated. The output gate o_t further controls whether the latest memory cell's value C_t will be transmitted to the final state H_t. With the above definitions, the ConvLSTM can be formulated as follows:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci}\, C_{t-1} + b_i),\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf}\, C_{t-1} + b_f),\\
C_t &= f_t\, C_{t-1} + i_t\, \tanh\left(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c\right),\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co}\, C_t + b_o),\\
H_t &= o_t\, \tanh\left(C_t\right),
\end{aligned}
\tag{2}
$$

where '*' denotes the convolution operator, '∘' denotes the Hadamard product, and σ is the sigmoid activation function. X_t and $\mathcal{H}_t$ are the input and output of the ConvLSTM at time step t (t indicates the tth frame in a CT image sequence, and slices will be referred to as frames in the sequel.), and i_t, f_t, and o_t indicate the input, forget and output gates, respectively. b_i, b_f, and b_o are the bias of the input gate, forget gate, and output gate. A memory cell C_t stores the historical information. All the gates i, f, o, memory cell C, hidden state $\mathcal{H}$ and the learnable weights W are 3D tensors. Input sequences ? are fed into a ConvLSTM block, which captures the long and short-term memory of sequences and contains both axial dimensional information, for use in implicitly learning axial dimensional dynamics and efficiently fusing axial dimensional features.

We define L_t as the output of the ConvLSTM layer at time step t. The output of the ConvLSTM layer is fed to the fully-connected (FC) layers, which transform the features into a space that makes the output easier to classify. The outputs of the FC layers are defined as O_t at time step t. Ideally, the longer the image sequence, and the more classification information ConvLSTM processes, the higher the confidence of classification. From this perspective, it is sufficient to use the output of the final time step for classification without further processing. However, in practice, due to differences in the distribution of lesions on different slices, there may be some useful information that has not been accumulated in the memory

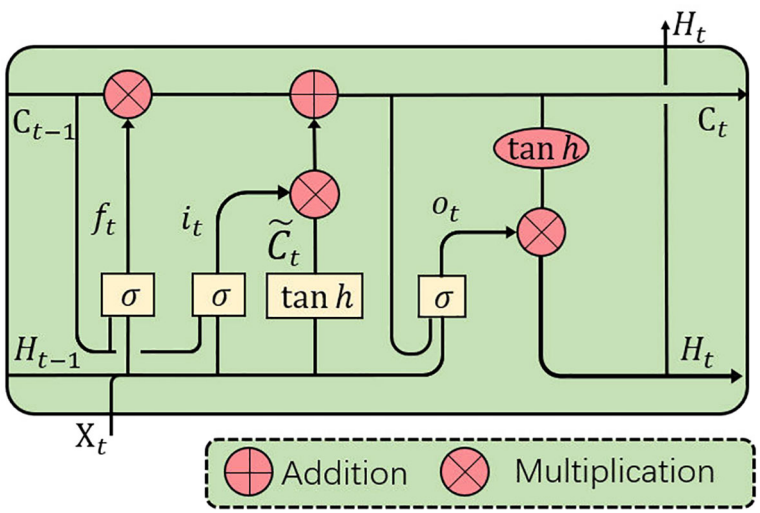

FIGURE 4 | ConvLSTM is utilized to implicitly learn axial dimensional dynamics and efficiently fuse axial dimensional features.

cell. In order to enchance the memory ablity of ConvLSTM for CT sequence at different slices and ensure that as much valid information from the previous slices as possible are encoded, we propose to use all the intermediate outputs of every time step as our feature for identification. A better ConvLSTM means that the longer the sequence it processes and the more comprehensive information it considers, the more confident it identifies the input. From this perspective, instead of minimizing the loss on the final time step, we define a new adaptive-weighted cross-entropy (ACE) loss to use all the intermediate outputs of every time step weighted by w_t:

$$\mathcal{L}_{ACE} = \frac{1}{n}\sum_{t=1}^{n}\sum_{p=1}^{P} - w_t[y_p \log (C_p(O_t))], \tag{3}$$

where C and p denote the classifier and classification label, respectively, and n denotes the number of images in a sequence. $C_p(O_t)$ indicates the classifier C, which correctly identifies the final output O at time step t, $y_p \in \{0, 1, 2\}$ are the label values; and P=3 denotes the total number of labels. Finally, w_t is the weight of each frame in a sequence. weight. We let each group of two weight items constitute the arithmetic sequence.

Since the importance of the information contained in different slices is different, it is not reasonable to use the equal weights. Due to the output of the final time step has taken into account all other previous slices , it contains the most information, and the further away from the last output, the less information it contains. Moreover, the number of slices in a sequence is a hyper-parameter, we adopt an adaptive weighting scheme. The output of the final time step should be assigned the maximum weight, and the farther away from the final time step, the smaller the weight. Specifically, we let each group of two weight items constitute the arithmetic sequence: the sum of which is 1. The first two items are taken as 0.01, namely, $w_1 = w_2 = 0.01$, and the subsequent weights can then be calculated according to the hyperparameter n and weight w_1. The ACE loss ensures that the features of the previous CT images in the sequence can be encoded into the later image.

3.2.3 Auxiliary Diagnosis With Attention Maps

Deciding which type (VP, BP, or normal) the entire volume belongs to based on the prediction results of the sequence is a critical step in auxiliary diagnosis. For higher confidence, in addition to using the model to obtain the classification result of the sequence, we also utilized the Grad-CAM (44) technology to generate attention maps of the sequence to assist the prediction. Grad-CAM is a method for producing visual interpretations for CNNs in the form of class-specific saliency maps. A saliency map, L_t^c, is produced for each image input based on the activation from k filters, A_{ij}^k, at the final convolutional layer. To make the method applicable to image sequences, the activations for all timesteps t in the sequence are considered (Eq. R1).

$$L_{ijt}^c = \sum_a w_{kt}^c A_{ijt}^k;\ w_{kt} = \frac{1}{Z}\sum_{ij}\frac{\partial F^c}{\partial A_{ijt}^k} \tag{R1}$$

where Z is a normalizing constant and F^c is the network output for the class c. i, j are pixel location of filter A^k. In the visualization examples shown in **Figure 5**, stronger class activation map (CAM) areas are indicated with lighter colors.

During the prediction phase, we can obtain the classification results of sequences belonging to a volume. In addition, we apply Grad-CAM to generate the response heat map of each sequence. A volume containing m sequences will be classified as viral pneumonia (VP) sample if it meets both of following criteria: (a) More sequences are classified as viral pneumonia (VP) then bacterial pneumonia (BP) in this volume; (b) There are two adjacent sequences with the category of viral pneumonia whose activation regions have an intersecting area of more than 50%. If the second criterion is not satisfied for VP, the bacterial type sequences are checked if there are two adjacent sequences with an intersecting area of more than 50% of the activation region, and if so the sample is classified as bacterial type, otherwise it is classified as normal. Notably, if there are the same number of VP and BP sequences, the one with the greater average sequence probability value is treated as the dominant category. The possibility output of the network dominates the classification on the volume level, and Attention Map is used as an auxiliary during the decision making.

3.3 Evaluation Metrics

We employ the commonly used metrics for multi-class classification to measure performance: e.g., weighted sensitivity (Sen, also known as recall), specificity (Spe), accuracy (Acc), and balanced accuracy (B-Acc, a.k.a. balanced classification rate). In order to reflect the tradeoff between sensitivity and specificity, and evaluate the quality of our classification results more reliably, a kappa analysis and F-measure ($F1$ score) are also provided following (45). These two measures are more robust than other percentage agreement measures, as they take into account the possibility of the agreement occurring by chance. The weighted sensitivity (Sen), specificity (Spe), accuracy (Acc), and balanced accuracy (B-Acc) are defined as:

$$Spe = \sum_{i=1}^{P} w_i \frac{TN_i}{TN_i + FP_i},\ Sen = \sum_{i=1}^{P} w_i \frac{TP_i}{TP_i + FN_i},$$

$$B - Acc = \frac{(Sen + Spe)}{2},\ Acc = \sum_{i=1}^{P} w_i \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i},$$

where TP_i indicates the number of true positives, TN_i—the number of true negatives, FP_i—the number of false positives, and FN_i—the number of false negatives for the i–*th* classification label; and w_i represents the percentage of images whose ground truth labels are i. The kappa values and F-measure ($F1$ score, a.k.a. Dice score) are defined as follows:

$$p_o = \sum_{i=1}^{P} w_i \frac{TP_i}{n}, p_e = \sum_{i=1}^{P} \frac{a_i * b_i}{n * n},\ Pre = \sum_{i=1}^{P} w_i \frac{TP_i}{TP_i + FP_i},$$

$$Kappa = \frac{p_o - p_e}{1 - p_e},\ F1 = 2 \cdot \frac{Pre \cdot Sen}{Pre + Sen},$$

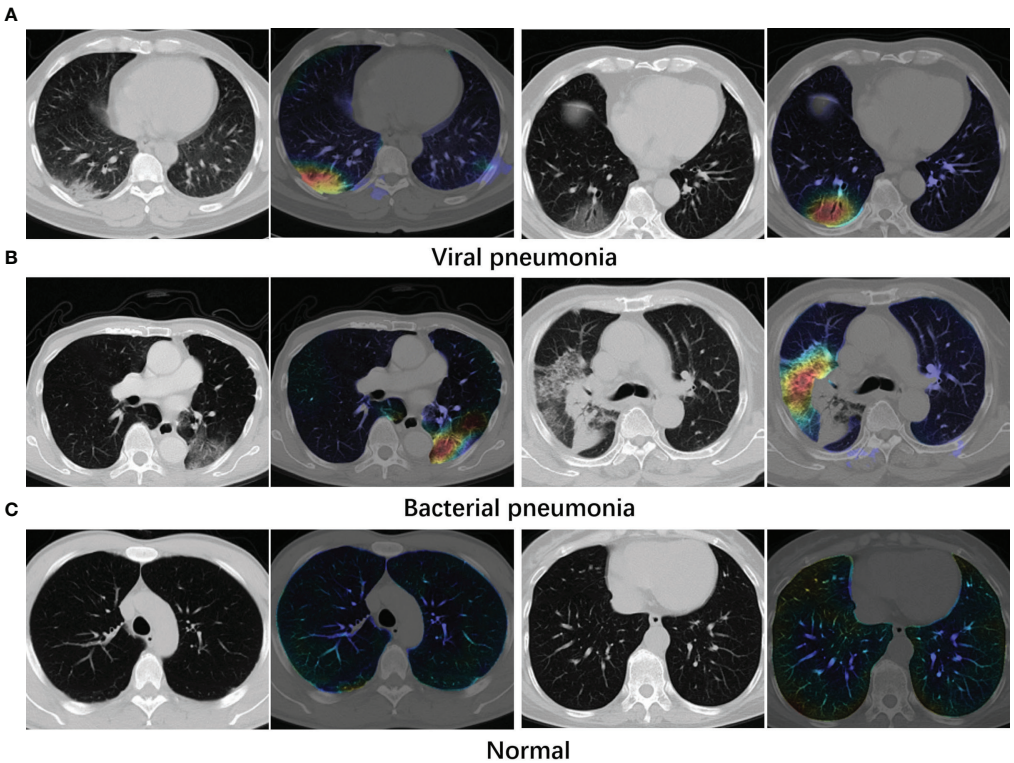

FIGURE 5 | Examples of attention maps obtained with Grad-CAM. **(A)** Viral pneumonia cases. **(B)** Bacterial pneumonia cases. **(C)** Normal cases. Lighter colors indicate the stronger response regions. From the maps, the infected regions receive greater attention.

where a_i denotes the true sample number of each class, b_i denotes the predicted sample number of each class, n denotes the total sample number, and P denotes the number of classes. Note that *kappa* values between 0.81 to 1.00 indicate almost perfect agreement, values between 0.61 and 0.80 exhibit substantial agreement, values of 0.41–0.60 exhibit moderate agreement and values less than 0.40 exhibit poor to fair agreement. The $F1$ score reaches its best value at 1 and worst at 0. We also present the ROC curves and the area under ROC curve (AUC) for VP against BP.

3.4 Implementation Details

The proposed method was implemented in the publicly available Pytorch library. The combination of CNN and ConvLSTM makes the model more complex. To accelerate convergence, we first trained a CNN classification network with a labeled 2D slice. After removing the FC layer, the encoder is used as the initialization parameter of SLP-Net. During the training phase of SLP-Net, CNN and ConvLSTM are jointly trained in an end-to-end manner using Adam optimizer. In practice, we found that CNN pre-training does speed up the convergence of the model, but has no effect on the final classification performance. The learning rate was gradually decreasing starting from 0.0001, and the momentum was set to 0.9. In addition, online data enhancement was employed to enlarge the training sequence data. The same data enhancement was used for all images in a sequence: we implemented data augmentation in a random way, including brightness, color, contrast, and sharpness transformation from 90 to 110%. We set a random seed from 1 to 4 for the enhancement.

4 RESULTS

4.1 Classification Performances

To compare the classification performance, we evaluated the detection ability of the model at both sequence and volume levels. All the existing pneumonia detection methods are accomplished using 2D CT slices or 3D volumes. To further verify whether the features containing both axial dimensional and spatial information captured by our model could benefit detection performance, we compared the proposed method to other classic classification models using 2D slices: AlexNet (46), VGG19 (47), InceptionV3 (48), ResNet34 (36), and Xception (49). Due to the lack of sufficient training data and the GPU memory constraint, we cannot apply 3D CNNs on complete CT volumes. In order to compare the proposed SLP-Net with other 3D deep learning architectures, we apply CT sequence data, which can be considered as 3D data, to train 3D models, including C3D (50), I3D (51), and S3D (52). We report the detection results for slice/sequence-level and case-level in **Table 2**. We applied a similar strategy to that in *Auxiliary*

TABLE 2 | Classification results for VP, BP and normal controls by different methods.

Method	Slice/Sequence-Level					Case-level				
	Kappa	F1	B-Acc	Sen	Spe	Kappa	F1	B-Acc	Sen	Spe
AlexNet	0.5207	0.5680	0.6872	0.6375	0.7370	0.6889	0.7381	0.8207	0.8274	0.8140
VGG19	0.6258	0.6574	0.7502	0.7152	0.7853	0.7709	0.8000	0.8571	0.8690	0.8601
ResNet34	0.6767	0.7100	0.7948	0.7783	0.8112	0.8489	0.8598	0.9045	0.9048	0.9043
InceptionV3	0.5177	0.6107	0.7156	0.6978	0.7333	0.7692	0.7976	0.8607	0.8631	0.8582
Xception	0.6802	0.6776	0.7688	0.7252	0.8124	0.8500	0.8631	0.9048	0.9107	0.9061
C3D	0.7382	0.7442	0.8232	0.8013	0.8450	0.8616	0.8729	0.9158	0.9183	0.9132
I3D	0.7437	0.7513	0.8403	0.8302	0.8132	0.8481	0.8952	0.9229	0.9167	0.9290
S3D	0.7525	0.7332	0.8106	0.7696	0.8517	0.8696	0.8796	0.9297	0.9133	0.9157
SLP-Net	**0.8280**	**0.8123**	**0.8665**	**0.8397**	**0.8934**	**0.9263**	**0.9291**	**0.9523**	**0.9524**	**0.9521**

The best performance of all the methods is highlighted in bold.

Diagnosis With Attention Maps section to determine the prediction result of a volume when using the 2D models. First, Grad-CAM was used to generate the activated maps of 2D slices, and binary activated maps can be obtained through thresholding. If five consecutive slices in a volume were predicted as indicative of viral pneumonia, and the intersection area of their activated area exceeds 50% of the union area, the volume was considered to be indicative of viral pneumonia.

As can be observed, the 3D networks achieve higher performances than the 2D networks, which confirms the importance of the combination of axial dimensional and spatial information for accurate detection results. At a slice/sequence level, our SLP-Net outperformed other methods in terms of *kappa*, *F*1 and *Sen* by a large margin, as well as achieving the best performance at a volume level. In addition, **Figure 6** shows the confusion matrices of VGG-19, Resnet34, Xception, C3D,

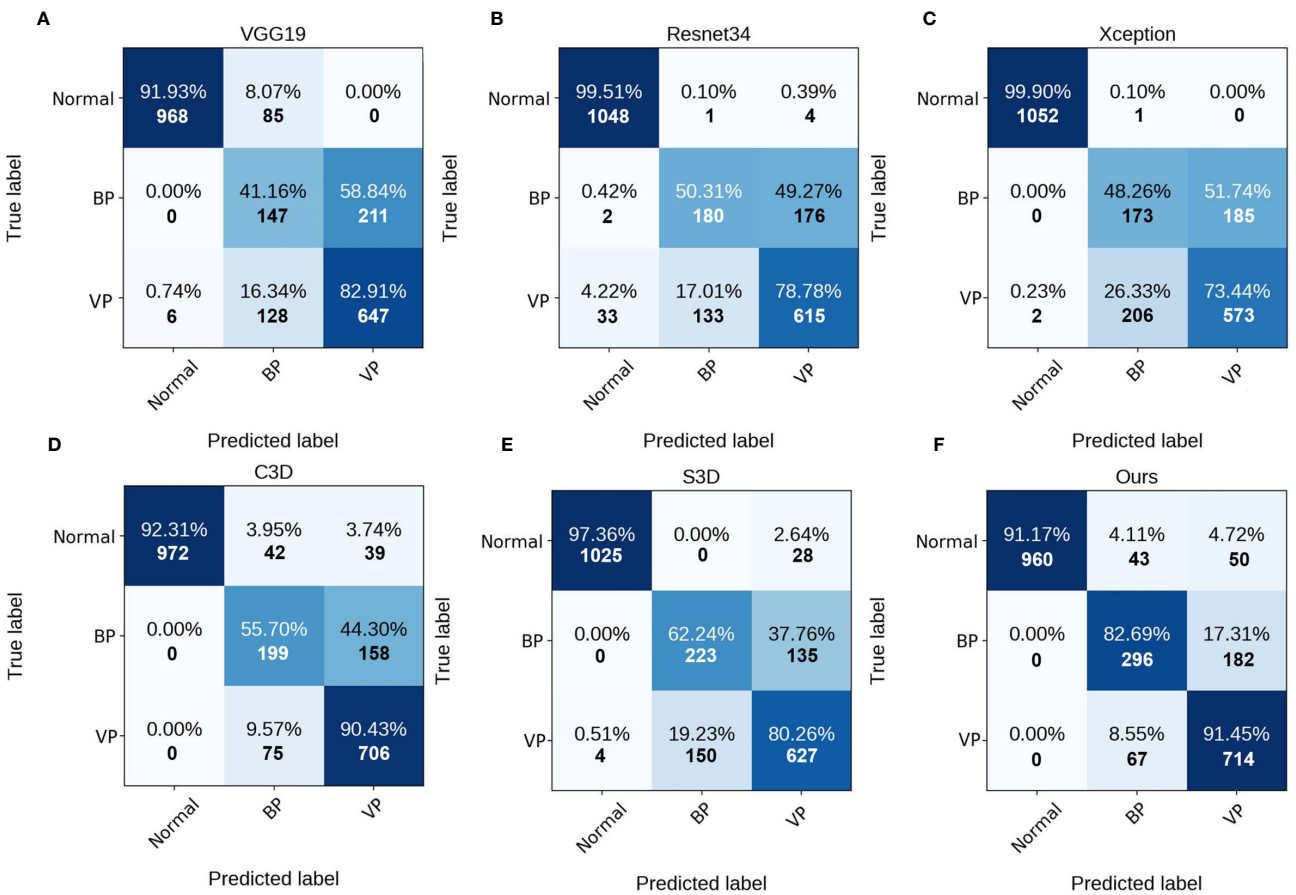

FIGURE 6 | Confusion matrices of the different methods at slice/sequence level. **(A–F)** are the results of VGG19, Resnet34, Xception, C3D, S3D and ours, respectively. The numbers in the confusion matrices denote the percentage (above) and number (below) of the predicted class.

S3D, and our method over the dataset. These results further indicate the superiority of the performance of our approach. As stated in the Introduction section, the difficulty of pneumonia diagnosis is the differentiation between BP and VP. Accordingly, we conducted experiments on the dataset contained VP and BP samples only. Results are shown in **Table 3** and **Figure 7**. We may observe that the proposed method again produces the best performance compared to the other methods. **Table 3** gathers all the performances of these models. The results show that the 3D-based method is generally better than the 2D-based method, mainly because the 3D input provides richer spatial information, which allows the model to learn and extract the subtle differences in the spatial distribution of different diseases, which is especially important for difficult samples with similar lesion appearance. In this regard, our proposed method not only utilizes the 3D information, but also explicitly focuses on the lesion area in the decision-making process through attention map, which makes the classification results more reliable. This idea can be applied to many different medical image-based classification tasks, since the similarity of lesion appearance is a problem in many scenarios.

Table 4 summarizes space and time cost of different methods. For fair comparison of inference time, we test all these models with PyTorch. Our SLP-Net had the best time efficiency and achieved smallest model size because it didn't use 3D convolution operations.

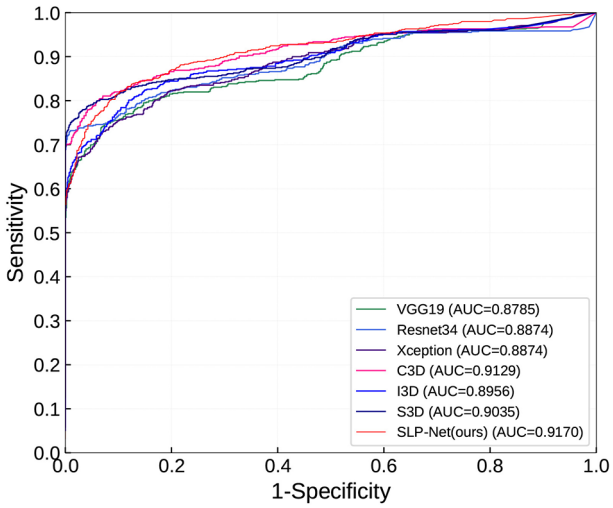

FIGURE 7 | ROC curves in classifying viral pneumonia (VP) and bacterial pneumonia (BP) of the compared models and the proposed method.

5 ANALYSIS AND DISCUSSION

5.1 Sensitivities to Hyperparameters

In **Table 5** we investigate the sequence settings, i.e., n and k, denote the number of slices per sequence and the number of overlapping slices between two sequences, respectively. By default we set $n = 10$ and $k = 5$. As can be observed, the performance was greatly affected by the value assigned to n. When n is set very small ($n = 5$), the *kappa* and $F1$ drop by the considerable margin of 3%, demonstrating that the more axial dimensional information ConvLSTM encodes, the more the model benefits. However, when n is greater than 15, the model performance will decline. This may be due to the fact that as the slice number increases, there will be less training data, resulting in the model not being fully trained. **Table 5** also shows that our result is impacted just marginally when k is within a scale of 5-7. The performance of the model mainly depends on the abundance of the information contained in the sequence, that is, the more information contained in the sequence, the better the classification performance. Compared to n = 10, the sequence provide less timing information when n = 5, so the classification performance will decrease. If n is too large (e.g., n = 20), the performance will decrease due to fewer training samples.

5.2 Ablation Study

Our SLP-Net employs three main components to form the classification framework: a sequence CNNs with multi-scale receptive fields, a ConvLSTM module, and a carefully designed ACE loss. In this subsection, we analyze and discuss the network under different scenarios to validate the performance of each key component of our model, and the results of different combinations of these modules are reported in **Figure 8**.

5.2.1 Effectiveness of ConvLSTM

To explore the contribution of the ConvLSTM, we use a ResNet50 pretrained on ImageNet as the backbone. As shown in **Figure 8**, a backbone + ConvLSTM + Res2Net method clearly

TABLE 3 | Comparison of different methods in classifying viral pneumonia (VP) and bacterial pneumonia (BP), at a slice/sequence level.

Method	Acc	Sen	Spe	AUC (*p*-value)
AlexNet	0.7327	0.8182	0.6650	0.8700 (p <0.001)
VGG19	0.7776	0.8197	0.7442	0.8785 (p <0.001)
ResNet34	0.7939	0.8305	0.7649	0.8874 (p <0.001)
InceptionV3	0.7429	0.8028	0.6955	0.8697 (p <0.001)
Xception	0.8170	0.7704	0.8438	0.8874 (p <0.001)
C3D	0.8218	**0.869**	0.7844	0.9129 (p <0.05)
I3D	0.8320	0.8274	0.8356	0.8956 (p <0.001)
S3D	0.8361	0.8413	0.8319	0.9035 (p <0.01)
SLP-Net	**0.8476**	0.8459	**0.8490**	**0.9170**

P-value is calculated by Delong's test.
The best performance of all the methods is highlighted in bold.

TABLE 4 | Model size and inference time of different methods.

	C3D	I3D	S3D	SLP-Net
Model size (MB)	39.2	48.7	42.3	34.4
Time (ms)	41.4	59.0	47.1	39.5

TABLE 5 | Effect of different settings of hyperparameter *n* and *k* on the results.

Method	kappa	F1	Acc	B-Acc	Sen	Spe
n = 5, k = 3	0.7376	0.7469	0.8546	0.8272	0.8096	0.8449
n = 10, k = 3	0.7652	0.7701	0.8684	0.8453	0.8307	0.8599
n = 10, k = 5	0.8241	0.8091	0.9012	0.8641	0.8471	0.8911
n = 10, k = 7	0.8280	0.8123	0.9034	0.8665	0.8397	0.8934
n = 15, k = 5	0.7453	0.7557	0.8577	0.8362	0.8231	0.8493
n = 20, k = 5	0.7329	0.7068	0.8577	0.7930	0.7450	0.8410

Here, n *and* k *denote the number of slices in the sequence and the number of overlapping slices between two sequences, respectively.*

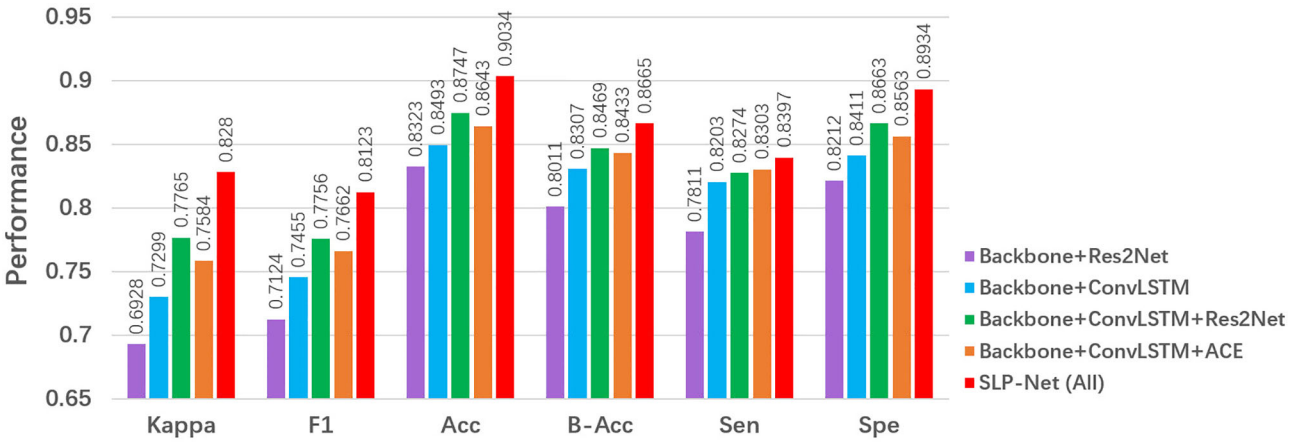

FIGURE 8 | Ablation studies of our SLP-Net.

TABLE 6 | Ablation study of Attention Map in classifying VP, BP, and Normal controls at the case-level.

Method	Kappa	F1	B-Acc	Sen	Spe
Without Attention Map	0.8731	0.9090	0.9167	0.9174	0.9160
With Attention Map	0.9263	0.9291	0.9523	0.9524	0.9521

outperformes the backbone + Res2Net, with improvement of about 6% in *F*1. This shows that the ConvLSTM is capable of extracting the axial dimensional and spatial information, thus memorizing the change in appearance that corresponds to axial dimensional information, and improving the performance in identifying and discriminating between VP and BP.

5.2.2 Effectiveness of the Res2Net module

We investigated the importance of the multi-scale sequence module, i.e., Res2Net. From **Figure 8**, we observe that a backbone + ConvLSTM + Res2Net model outperformed the backbone model in terms of major metrics, i.e., *kappa* and *F*1. This suggests that introducing the Res2Net module enables the encoder to capture more discriminative features to accurately differentiate VP from BP.

5.2.3 Effectiveness of ACE

Finally, we investigate the importance of the ACE loss. From the results in **Figure 8**, it may clearly be observed that the ACE Loss effectively improves the classification performance in our model. One possible reason is that, with the ACE loss, the ConvLSTM explores the axial dimensional dynamics of appearance features in CT sequences, and these features are further aggregated for classification purposes.

5.2.4 Effectiveness of Attention Maps

To investigate the contribution of the Attention Map, we added an additional experiment—case-level classification without Attention Map. Specifically, sequence-level classification results were first obtained using SLP-Net, and if there were VP or BP sequences in a volume, the type with more number is used as the category of the whole volume. If it does not contain VP and BP sequence, it is classified as normal. Notably, if there are the same number of VP and BP sequences, the one with the greater average sequence probability value is treated as the dominant category. **Table 6** shows the result, where SLP-Net with Attention Map as auxiliary achieves better performance than without Attention Map. This demonstrates that with the aid of attention map, the distribution of lesions can be considered simultaneously in the decision-making process, thus improving the performance of case-level classification.

6 DISCUSSION AND CONCLUSIONS

6.1 Limitations

Although our method achieves better results in the pneumonia classification task compared to other methods, this work still has some limitations. Firstly, we used the multiscale feature technique Res2Net in the feature extraction part, but did not further explore the hyperparameter settings in it, and although we believe that careful selection of hyperparameters may further improve the classification performance, no additional experiments were conducted in this work to compare the impact of different hyperparameters since this is not the focus of our work. Secondly, the model is not evaluated on an external dataset. To our knowledge, there are no publicly available 3D CT datasets for different types of pneumonia classification tasks, and it is difficult to collect compliant data from multiple centers due to various conditions. We intend to evaluated the performance of our model on external datasets in the future.

6.2 Conclusion

Hospitals are beginning to use CT imaging in the diagnosis of viral pneumonia, and it is vital to improve the sensitivity of diagnostic methods so as to reduce the incidence of false negatives. AI-empowered image acquisition workflows are effective, and may also aid in protecting clinicians from viral pneumonia (e.g., COVID-19) infection. Although several effective AI-based COVID-19 diagnosis or lesion segmentation methods have been introduced recently, automated differentiation of viral pneumonia from other types of pneumonia is still a challenging task. The motivation of this study was to employ AI techniques to alleviate the problem posed by the fact that even radiologists are hard pressed to distinguish VP from BP, as they share very similar presentations of infection lesion characteristics in CT images.

In this paper, we have proposed a novel viral pneumonia detection network, named SLP-Net. By contrast with previous 2D slice-based or 3D volume-based methods, we considered continuous CT images as time sequences. Our model first utilized the sequence CNNs with multi-scale receptive fields to extract a sequence of higher-level representations. The feature sequences were then fed into a ConvLSTM to capture axial dimensional features. Finally, in order to ensure that as many valid features from previous slice as possible are encoded into the later CT slices, a novel ACE loss was proposed to optimize the output of the SLP-Net. Furthermore, during the prediction phase, we used sequence attention maps for auxiliary classification to predict each volume, which can enhance the confidence level of the results. Overall, in order to accurately distinguish VP from BP and normal subjects, we used the sequence CNNs with multi-scale receptive fields to extract more differentiating features, and then applied a ConvLSTM to capture axial dimensional features of the CT sequence, thereby obtaining features containing both axial dimensional and spatial information. The superior evaluation performance achieved in the classification experiments demonstrate the ability of our model in the differential diagnosis of VP, BP and normal cases. Although we only evaluated our method on the CT dataset of pneumonia, it can be adapted to any other 3D medical image classification problems, such as lung cancer imaging analysis, and the identification of Alzheimer's disease. In future work we will further validate our models on even larger datasets, and seek its implementation in real clinical settings.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Hwa Mei Hospital, University of Chinese Academy of Sciences. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JH, JX, and RL were involved in data analysis and interpretation, and drafting and revising the manuscript. HH, YM, KY, RRL, YLZ, and JJZ were involved in data analysis and interpretation. JL, JFZ, and YTZ were involved in study conceptualization, supervision, revising the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China (LZ19F010001, and LQ20F030002), in part by the Key Project of Ningbo Public Welfare Science and Technology (2021S107), and in part by the Youth Innovation Promotion Association CAS (2021298).

REFERENCES

1. Wang C, Horby PW, Hayden FG, Gao GF. A Novel Coronavirus Outbreak of Global Health Concern. *Lancet* (2020) 395:470–3. doi: 10.1016/S0140-6736(20)30185-9
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical Features of Patients Infected With 2019 Novel Coronavirus in Wuhan, China. *Lancet* (2020) 395:497–506. doi: 10.1016/S0140-6736(20)30183-5
3. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology* (2020) 296:E32–40. doi: 10.1148/radiol.2020200642
4. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology* (2020) 296:E115–7. doi: 10.1148/radiol.2020200432
5. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest Ct for Covid-19: Comparison to Rt-Pcr. *Radiology* (2020) 296:E115–7. doi: 10.1148/radiol.2020200432
6. Pan F, Ye T, Sun P, Gui S, Liang B, Li L, et al. Time Course of Lung Changes at Chest CT During Recovery From Coronavirus Disease 2019 (COVID-19). *Radiology* (2020) 295:715–21. doi: 10.1148/radiol.2020200370
7. McLuckie A. *Respiratory Disease and its Management*. London, UK: Springer Science & Business Media (2009).
8. Miller WT, Mickus TJ, Barbosa JE, Mullin C, Van VM, Shiley K, et al. CT of Viral Lower Respiratory Tract Infections in Adults: Comparison Among Viral Organisms and Between Viral and Bacterial Infections. *Am J Roentgenology* (2011) 197:1088–95. doi: 10.2214/AJR.11.6501
9. Jae YP, Rosemary F, Richard S, Neil S, Nicholas J. *The Accuracy of Chest Ct in the Diagnosis of Covid-19: An Umbrella Review. Website* (2021). Available at: https://www.cebm.net/covid-19/the-accuracy-of-chest-ct-in-the-diagnosis-of-covid-19-an-umbrella-review/.
10. Nambu A, Ozawa K, Kobayashi N, Tago M. Imaging of Community-Acquired Pneumonia: Roles of Imaging Examinations, Imaging Diagnosis of Specific Pathogens and Discrimination From Noninfectious Diseases. *World J Radiol* (2014) 6:779. doi: 10.4329/wjr.v6.i10.779
11. Wootton D, Feldman C. The Diagnosis of Pneumonia Requires a Chest Radiograph (X-Ray)—Yes, No or Sometimes? *Pneumonia* (2014) 5:1–7. doi: 10.15172/pneu.2014.5/464
12. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for Covid-19. *IEEE Rev Biomed Eng* (2020) 14:4–15. doi: 10.1109/RBME.2020.2987975
13. Dong D, Tang Z, Wang S, Hui H, Gong L, Lu Y, et al. The Role of Imaging in the Detection and Management of Covid-19: A Review. *IEEE Rev Biomed Eng* (2020) 14:16–29. doi: 10.1109/RBME.2020.2990959
14. Ouyang X, Huo J, Xia L, Shan F, Liu J, Mo Z, et al. Dual-Sampling Attention Network for Diagnosis of Covid-19 From Community Acquired Pneumonia. *IEEE Trans Med Imaging* (2020) 39:2595–605. doi: 10.1109/TMI.2020.2995508
15. Shi F, Xia L, Shan F, Song B, Wu D, Wei Y, et al. Large-Scale Screening to Distinguish Between Covid-19 and Community-Acquired Pneumonia Using Infection Size-Aware Classification. *Phys Med Biol* (2021) 66:065031. doi: 10.1088/1361-6560/abe838
16. Fan D, Zhou T, Ji G, Zhou Y, Chen G, Fu H, et al. Inf-Net: Automatic Covid-19 Lung Infection Segmentation From Ct Images. *IEEE Trans Med Imaging* (2020) 39:2626–37. doi: 10.1109/TMI.2020.2996645
17. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *cell* (2018) 172:1122–31. doi: 10.1016/j.cell.2018.02.010
18. Rajaraman S, Candemir S, Kim I, Thoma G, Antani S. Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. *Appl Sci* (2018) 8:1715. doi: 10.3390/app8101715
19. Han Z, Wei B, Hong Y, Li T, Cong J, Zhu X, et al. Accurate Screening of Covid-19 Using Attention-Based Deep 3d Multiple Instance Learning. *IEEE Trans Med Imaging* (2020) 39:2584–94. doi: 10.1109/TMI.2020.2996256
20. Kohno N, Ikezoe J, Johkoh T, Takeuchi N, Tomiyama N, Kido S, et al. Focal Organizing Pneumonia: CT Appearance. *Radiology* (1993) 189:119–23. doi: 10.1148/radiology.189.1.8372180
21. Xie W, Jacobs C, Charbonnier JP, Van Ginneken B. Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in Ct Scans. *IEEE Trans Med Imaging* (2020) 39:2664–75. doi: 10.1109/TMI.2020.2995108
22. Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, et al. A Noise-Robust Framework for Automatic Segmentation of Covid-19 Pneumonia Lesions From Ct Images. *IEEE Trans Med Imaging* (2020) 39:2653–63. doi: 10.1109/TMI.2020.3000314
23. Narin A, Kaya C, Pamuk Z. Automatic Detection of Coronavirus Disease (Covid-19) Using X-Ray Images and Deep Convolutional Neural Networks. *Pattern Anal Appl* (2021) 24:1207–20. doi: 10.1007/s10044-021-00984-y
24. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. AI Augmentation of Radiologist Performance in Distinguishing COVID-19 From Pneumonia of Other Etiology on Chest CT. *Radiology* (2020) 296:E156–65. doi: 10.1148/radiol.2020201491
25. Barstugan M, Ozkaya U, Ozturk S. *Coronavirus (Covid-19) Classification Using Ct Images by Machine Learning Methods*. Konya, Turkey: arXiv preprint arXiv:2003.09424 (2020).
26. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Artificial Intelligence Distinguishes Covid-19 From Community Acquired Pneumonia on Chest Ct. *Radiology* (2020) 0:200905. doi: 10.1148/radiol.2020200905
27. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee J, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Network Open* (2019) 2:e191095–e191095. doi: 10.1001/jamanetworkopen.2019.1095
28. Wang X, Deng X, Fu Q, Zhou Q, Feng J, Ma H, et al. A Weakly-Supervised Framework for Covid-19 Classification and Lesion Localization From Chest Ct. *IEEE Trans Med Imaging* (2020) 39:2615–25. doi: 10.1109/TMI.2020.2995965
29. Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, et al. Deep Learning for Classification and Localization of Covid-19 Markers in Point-of-Care Lung Ultrasound. *IEEE Trans Med Imaging* (2020) 39:2676–87. doi: 10.1109/TMI.2020.2994459
30. Wang J, Bao Y, Wen Y, Lu H, Luo H, Xiang Y, et al. Prior-Attention Residual Learning for More Discriminative Covid-19 Screening in Ct Images. *IEEE Trans Med Imaging* (2020) 39:2572–83. doi: 10.1109/TMI.2020.2994908
31. Oh Y, Park S, Ye JC. Deep Learning Covid-19 Features on Cxr Using Limited Training Data Sets. *IEEE Trans Med Imaging* (2020) 39:2688–700. doi: 10.1109/TMI.2020.2993291
32. Alom MZ, Rahman M, Nasrin MS, Taha TM, Asari VK. *Covid_mtnet: Covid-19 Detection With Multi-Task Deep Learning Approaches*. Dayton, OH, USA: arXiv preprint arXiv:2004.03747 (2020).
33. Özkaya U, Öztürk Ş, Barstugan M. Coronavirus (Covid-19) Classification Using Deep Features Fusion and Ranking Technique. In: *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*. Konya, Turkey: Springer (2020). p. 281–95.
34. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (Covid-19) With Ct Images. *IEEE/ACM Trans Comput Biol Bioinf* (2021) 34:102–6. doi: 10.1109/TCBB.2021.3065361
35. Kang H, Xia L, Yan F, Wan Z, Shi F, Yuan H, et al. Diagnosis of Coronavirus Disease 2019 (Covid-19) With Structured Latent Multi-View Representation Learning. *IEEE Trans Med Imaging* (2020) 39:2606–14. doi: 10.1109/TMI.2020.2992546
36. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Las Vegas, NV, USA: CVPR (2016). p. 770–8.
37. Chen CF, Fan Q, Mallinar N, Sercu T, Feris R. *Big-Little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition*. New Orleans, USA: arXiv preprint arXiv:1807.03848 (2018).
38. Chen Y, Fan H, Xu B, Yan Z, Kalantidis Y, Rohrbach M, et al. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks With Octave Convolution. *CVPR* (2019), 3435–44. doi: 10.1109/ICCV.2019.00353
39. Cheng B, Xiao R, Wang J, Huang T, Zhang L. *High Frequency Residual Learning for Multi-Scale Image Classification*. Cardiff, Wales: arXiv preprint arXiv:1905.02649 (2019).
40. Gao S, Cheng M, Zhao K, Zhang X, Yang M, Torr P. Res2net: A New Multi-Scale Backbone Architecture. *IEEE Trans Pattern Anal Mach Intell* (2021) 43:652–62. doi: 10.1109/TPAMI.2019.2938758
41. Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput* (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

42. Song H, Wang W, Zhao S, Shen J, Lam KM. *Pyramid Dilated Deeper Convlstm for Video Salient Object Detection*. Munich, Germany: ECCV (2018) p. 715–31.
43. Shi X, Chen Z, Wang H, Dit-Yan, Yeung, Wai-kin W, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In: *International Conference on Neural Information Processing Systems*. Montreal, Canada: NIPS (2015) p. 802–C810.
44. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. *Grad-Cam: Visual Explanations From Deep Networks via Gradient-Based Localization*. Venice, Italy: ICCV (2017) p. 618–26.
45. Zhu J, Chen N, Xing EP. *Infinite Latent Svm for Classification and Multi-Task Learning*. Granada, Spain: NIPS (2011) p. 1620–8.
46. Krizhevsky A, Sutskever I, Hinton GE. *Imagenet Classification With Deep Convolutional Neural Networks*. Nevada, USA: NIPS (2012) p. 1097–105.
47. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. San Diego, California, USA: arXiv preprint arXiv:1409.1556 (2014).
48. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. *Rethinking the Inception Architecture for Computer Vision*. Las Vegas, NV, USA: CVPR (IEEE (2016) p. 2818–26.
49. Chollet F. *Xception: Deep Learning With Depthwise Separable Convolutions*. Honolulu, HI, USA: CVPR (2017) p. 1251–8.
50. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. *Learning Spatiotemporal Features With 3d Convolutional Networks*. Santiago, Chile: ICCV (2015) p. 4489–97.
51. Carreira J, Zisserman A. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. Honolulu, HI, USA: CVPR (2017) p. 6299–308.
52. Xie S, Sun C, Huang J, Tu Z, Murphy K. *Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification*. Munich, Germany: ECCV (2018) p. 305–21.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.